TOXICOLOGIC PATHOLOGY, vol 30, no 1, pp 15–27, 2002

Toxicogenomics, Drug Discovery, and the Pathologist

GARY A. BOORMAN,[1] STEVEN P. ANDERSON,[2] WARREN M. CASEY,[2] ROGER H. BROWN,[2] LYNN M. CROSBY,[3] K. GOTTSCHALK,[2] MARILYN EASTON,[2] HONG NI,[2] AND KEVIN T. MORGAN[2]

[1]*Laboratory for Experimental Pathology, Environmental Toxicology Program, NIEHS, PO Box 12233, 111 TW Alexander Drive, Research Triangle Park, NC 27709*
[2]*GlaxoSmithKline (GSK), 5 Moore Drive, Research Triangle Park, NC 27709, and*
[3]*US Environmental Protection Agency (EPA), NHEERL, Research Triangle Park, NC 27711*

ABSTRACT

The field of toxicogenomics, which currently focuses on the application of large-scale differential gene expression (DGE) data to toxicology, is starting to influence drug discovery and development in the pharmaceutical industry. Toxicological pathologists, who play key roles in the development of therapeutic agents, have much to contribute to DGE studies, especially in the experimental design and interpretation phases. The intelligent application of DGE to drug discovery can reveal the potential for both desired (therapeutic) and undesired (toxic) responses. The pathologist's understanding of anatomic, physiologic, biochemical, immune, and other underlying factors that drive mechanisms of tissue responses to noxious agents turns a bewildering array of gene expression data into focused research programs. The latter process is critical for the successful application of DGE to toxicology. Pattern recognition is a useful first step, but mechanistically based DGE interpretation is where the long-term future of these new technologies lies. Pathologists trained to carry out such interpretations will become important members of the research teams needed to successfully apply these technologies to drug discovery and safety assessment. As a pathologist using DGE, you will need to learn to read DGE data in the same way you learned to read glass slides, patiently and with a desire to learn and, later, to teach. In return, you will gain a greater depth of understanding of cell and tissue function, both in health and disease.

Keywords. Differential gene expression; genomics; proteomics; rodent studies; pathology; liver; research teams; microarray; toxicology; safety assessment.

INTRODUCTION

Rapid progress in genome sequencing and in the development of platforms to assess gene expression, protein expression and genetic polymorphisms (1, 5, 7, 15, 38, 40) has made these tools accessible to many research teams. Genomic technology promises to revolutionize research in drug discovery and toxicology (25, 33, 38). The power of large-scale differential gene expression (DGE) is that mRNA levels in cells can be obtained for thousands of genes in a single experiment (42). The technology currently requires considerable technical expertise, however. This has resulted in a team approach with scientists of diverse backgrounds working together to conduct toxicogenomic research. Pathologists, who have traditionally played a role in drug discovery and development, are now needed to work in teams using rapidly evolving technologies generating massive amounts of data.

The amount of data that will be available to investigators is unparalleled. For instance, GenBank, a public database, contained over 3.8 million sequence records at the end of the year 2000 and this database doubles in size approximately every year (24). The goal is to use this wealth of data to determine how different cellular components work together in health and disease (7, 31). The data may reveal novel drug targets, surrogate markers of efficacy or toxicity, or clues as to mechanism of toxicity. Traditionally, research has been on a 1-gene-1-protein at a time basis (the vertical approach) (49). In the genome era, horizontal investigations involve functional characterization of a large portion of the genes in a genome using single high-throughput tools. The horizontal and vertical approaches are complementary. Horizontal approaches offer the advantage of global analyses but do not provide conclusive answers. The vertical approach more appropriate to investigation of specific hypotheses, lacks efficiency, but can often answer questions raised by horizontal studies.

Structural genomics is directed towards understanding of the physical makeup of genomes, while functional genomics studies the function of the genes and gene products. Toxicogenomics is a subdiscipline of functional genomics using genomic tools to evaluate toxicity caused by pharmaceutical or environmental chemicals. DGE is already being used to complement histopathology and clinical chemistry in assessing toxicity. Interpreting mRNA data, in relation to whole-animal physiology, is to truly practice molecular pathology. As pathologists gain experience in this new discipline, toxicogenomics is revealed to be a familiar approach to investigative pathology, but using new and potentially powerful tools.

Address correspondence to: Gary A Boorman, Laboratory of Experimental Pathology, ETP, NIEHS, PO Box 12233, MD B3-08, 111 TW Alexander Drive, Research Triangle Park, North Carolina 27709.

0192-6233/02$3.00+$0.00

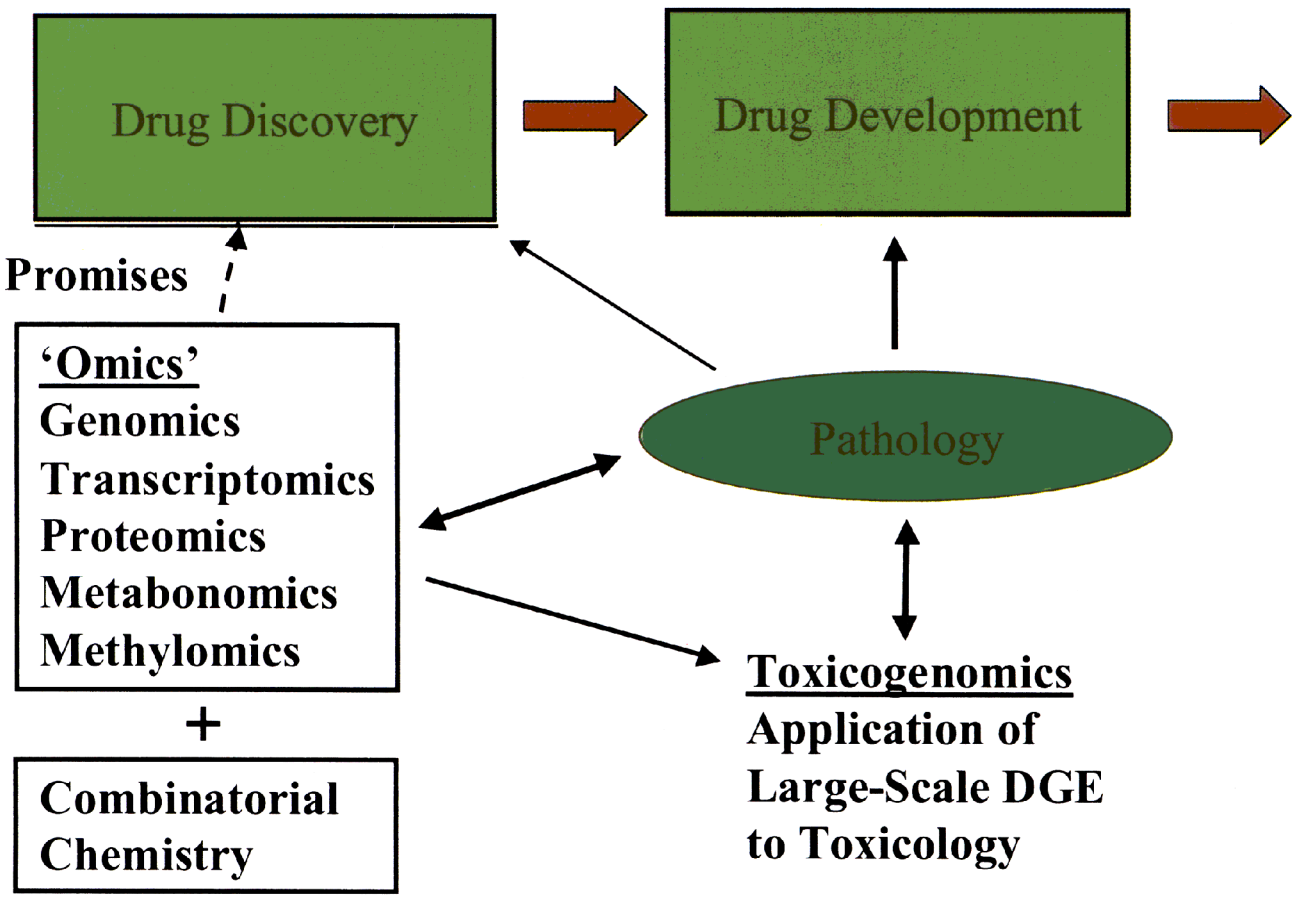


FIGURE 1.—Simplified diagram of the complex interplay of the flow of drug discovery and development, new technologies and the role of pathologists.

Although the information revolution may appear daunting, toxicogenomics has many features in common with pathology. Learning to function in the toxicogenomics team requires an understanding of the strengths and limitations of these new and evolving research techniques. Pathologists engaged in this learning process can play an essential role in integrating knowledge of cell biology, chemical toxicity, whole animal physiology, species-specific diseases, and other factors that impact interpretation of data derived from differential gene expression (DGE) studies. Pathologists and histotechnologists can address basic anatomical and physiological questions and monitor tissue selection quality. More importantly, pathologists provide detailed morphologic correlates with DGE data. Such correlates are considered critical by the authors.

In this article, we provide a brief introduction to gene expression array technology. This is followed by a description of one author's (KTM) approach to DGE data interpretation. A more detailed overview of DGE technology is provided as an appendix for those interested in delving deeper into this fascinating revolution in biology.

PATHOLOGISTS IN DRUG DISCOVERY AND THE 'OMICS' REVOLUTION

In a review of the process of drug discovery and development, Charles Smith (45) indicated clear roles for pathologists in later stages of drug discovery and in drug development. Often the major role of pathologists is in detecting or predicting untoward responses to new therapeutic agents. However, toxicologic pathologists are also appropriately members of research teams in the early stages of drug discovery. The 'omics' revolution provides an opportunity for pathologists to increase their contribution to drug discovery and development. The new technologies are radically changing biological research and providing an opportunity for pathologists to shape the development of toxicogenomics (Figure 1).

Role in Study Design

The role of the pathologist begins with study design. The pathologist is aware of the effects of feeding patterns, light, circadian rhythms, cage effects, and other stresses on the animals that affect gene expression. The experimental design should minimize these potential confounders. The pathologist also understands the substructure and complexity of the tissues under study. Tissues are not uniform (only 80% of the cells in the liver are hepatocytes and hepatocytes are not uniform throughout the lobule). Therefore some thought needs to be given to what tissues and what portions of tissues need to be collected. The collection of specific areas of tissue for DGE and for morphology should be part of the experimental design. For example, one would expect different DGE patterns from sampling whole kidney versus cortex, or renal medulla following exposure to a renal toxin.

Role in Tissue Sampling and Morphology

A complete necropsy with histopathologic evaluation of representative samples of key organ systems provides the morphological anchor and secondary corroboration of specific gene expression changes. Electron microscopy and immunohistochemistry can provide more detailed evidence of cellular changes to place DGE in perspective. In situ hybidization may identify specific cell types expressing target genes and allows detection of nonspecific hybridization that could be misinterpreted. Laser capture microdissection adds another tool to correctly identify and assess DGE patterns of specific cells (9).

Common morphological endpoints can be correlated with DGE. Apoptotic and mitotic cells are easily quantitated on H & E stained slides and would be expected to correlate with gene pathways that control these processes. Specialized stains such as TUNEL (terminal uridine deoxynucleotidyl transferase nick end labeling) gels demonstrating DNA "laddering" or the staining of annexin V are also markers for apoptosis. PCNA and BrdU stains provide quantitative measures of cell proliferation. Antibody-specific staining for proteins can provide morphological conformation for DGE-identified pathways related to oxidative stress or DNA repair. An anti–heme-oxygenase 1 (HO-1) immunohistochemical stain was used to visualize the protein product of the upregulated HO-1 gene (Figure 2) in cells treated with potassium bromate (13).

In vitro systems also benefit from morphologic examination. Measurement of LDH, trypan blue uptake and various measurements of ATP or NADH levels by oxidation/reduction of specific dyes provide only quantitation of cell viability. Histopathology of the cells can provide additional knowledge to aid in the interpretation of DGE data.

Role in Data Interpretation

Pathologists are in a unique position to interpret the underlying processes leading to complex DGE patterns following exposure to xenobiotics. Some xenobiotics induce their own metabolism, disturb homeostatic pathways and cause toxicity, all of which have the potential to influence DGE. Shifts in cellular populations may influence DGE patterns. For example, following chemical exposure, the liver response may include a mixture of both living and dead cells as well as inflammatory infiltrates. The RNA from exposed animals reflects this cellular mixture whereas controls lack the infiltrate. Another example is the comparison of colon cancer samples with normal colon that revealed more connective tissue and smooth muscle in the control tissue (37). Therefore, more highly expressed transcripts in normal tissue readily identified as of smooth muscle or connective tissue origin were excluded from the DGE analysis (37). Knowledge of the morphology and composition of control and treated tissues helps guide the DGE analysis.

The pathologist benefits from the investment in time in understanding basic intermediary metabolism and biochemistry, as well as the molecular biology of transcriptional controls. This basic information must then be updated with the constantly growing body of literature that surrounds each gene. Gene profiles often represent highly integrated cellular pathways. Gene analysis is more informative when the analytical approach includes gene linkages or cellular pathways. The reward for this integrative synthesis of morphology, physiology, pathogenesis, molecular biology, and biochemistry will be enormous.

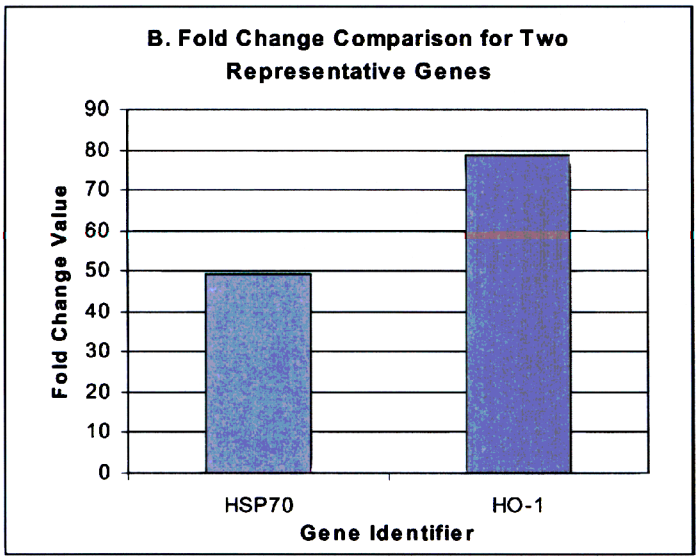
The authors' position is that toxicologic pathologists are ideally suited to undertake interpretation of DGE data. We must undertake this challenge energetically if we are to influence the outcome of the 'omics' revolution with respect to its impact on toxicology. It is hoped that toxicologic pathologists will also help to keep the "Bio" in Bioinformatics through positive influences on the evolution of this new and critical discipline that is currently being predominately influenced by statisticians and mathematicians. Learning to read DGE array data is a great place to start.

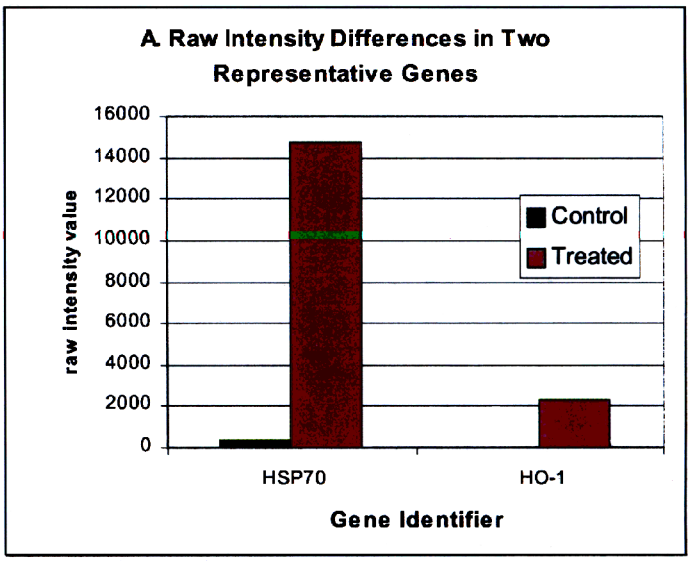
Gene Expression Array Data

Gene expression microarray technology uses the concept that mRNA is a complementary copy of the DNA coding region of its respective gene and will bind to complementary strands of DNA (42). DNA sequences from hundreds or thousands of genes can be attached to solid media such as glass or plastic slides. cDNA is made from the mRNA from treated or control cells and labeled with different fluorescent/radioactive markers. The cDNAs are allowed to competitively hybridize to the DNA or oligonucleotide probes on the slides. Using lasers that excite the specific dyes and detectors that convert the light to electrical signals in the case of fluorescent markers, one scans the slides for signals from thousands of genes. The mRNA expression levels for each of these genes is compared between the treated and controls (42). Other technologies, including those used by the authors, are comprised of nylon membranes with cDNA probes bound to the membrane as discrete dots that are hybridized to radioactive (generally ^{33}P or ^{32}P) cDNA copies of the mRNA population being investigated (Figure 3).

Gene expression often refers to the amount of the respective mRNA in cells, but downstream protein activity, in the majority of cases, represents the functional aspects of gene expression. Gene expression is regulated at several levels including (1) mRNA transcription, (2) processing, (3) transport, (4) degradation, (5) protein translation, and (6) protein activity. Gene expression provides a snapshot of the relative mRNA abundance's after steps 1 through 4. Western blots and techniques for assessing the posttranslational activity of a protein, such as electrophoretic mobility shift assays, complete the picture. In some cases, mRNA expression is a surrogate marker of gene function, because many genes are regulated at least in part at the level of transcription. It is critical when interpreting DGE data to account for the dynamics of these transcripts. Each transcript has a different half-life controlled by many factors including the mRNA sequence (2) and by a number of mRNA binding proteins. Translation is also regulated by many different factors (34) including the availability of the translational machinery (ribosomes) and necessary substrates (eg, transfer RNA and specific amino acids). An array of metabolic pathways provides both direct and indirect feedback to both mRNA transcription and translation.

There are limitations of DGE profiling studies in toxicology (15). First, many toxicants and drugs initiate toxicity by binding to proteins or altering macromolecules, not by directly altering gene expression. Moreover, multiple cellular signaling pathways alter the expression of the same gene products, making it difficult to identify the affected pathway





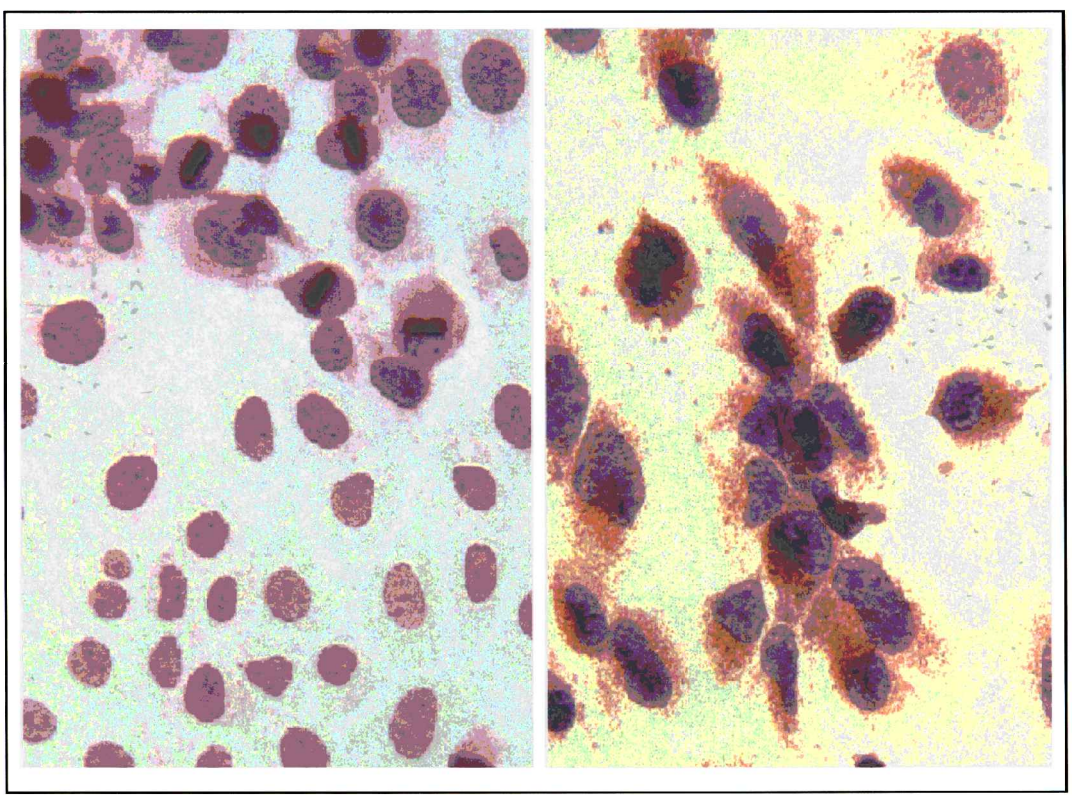


FIGURE 2.—A & B: Comparison of gene expression data for HO-1 and HSP70 genes to contrast raw intensity differences against expression of results as fold change (ratio). The normally low level of HO-1 expression exaggerates the treatment response when viewed only as a ratio. The response by HSP70 exhibits a higher overall signal intensity following treatment, and thus probably expression levels, than HO-1. The latter observation is not evident from ratios alone. C: HO-1 immunostain of control rat peritoneal mesothelial cells, D: HO-1 immunostain of cells treated with 3 mM $KBrO_3$ for 24 h. The use of antibody staining confirmed the differential gene expression-identified change in HO-1 expression vs control.

from DGE alone. Finally, most xenobiotics act through multiple mechanisms that depend on dose, timing, and duration of exposure. Variations in age, gender, temperature, light, diet, feeding, and hormonal status also affects DGE. DGE results must be integrated with the effects of toxicity within the context of the whole organism, but such interpretation is at the heart of pathology.

ANALYSIS OF DGE DATA SETS

Confirmation of DGE

Comparisons between treated and control animals for thousands of genes requires sophisticated analysis (52). The field of bioinformatics is evolving to provide the mathematical and statistical support for the evolving field of genomics

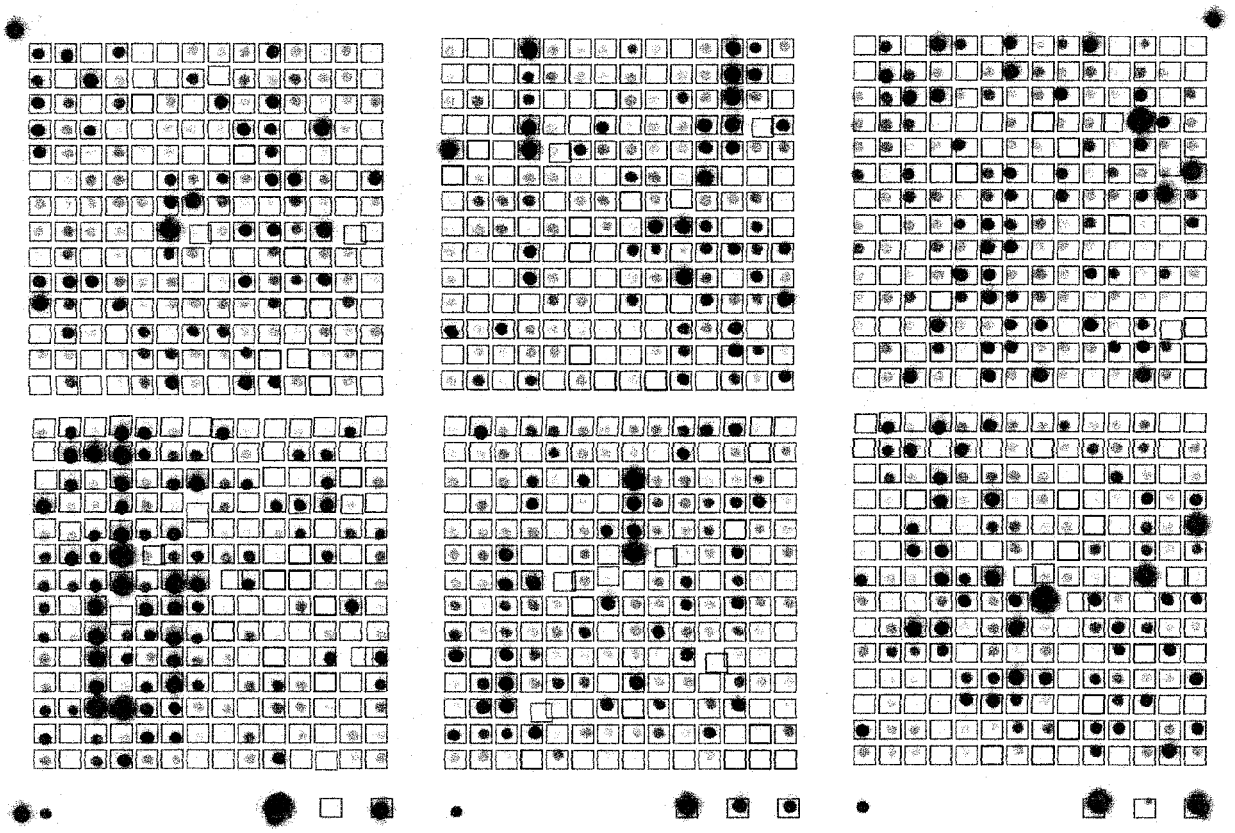


FIGURE 3.—Nylon DGE arrays, spotted for 1200 genes and hybridized against cDNA prepared from rat adipocytes and labeled with ^{33}P. Intensity of black spot indicates signal strength. Overlay generated by Clontech AtlasImage software during array alignment.

(24, 41). DGE results may include both false positive and false negatives, the number of which depends on the number of genes being analyzed, the statistical methods, the level of significance, the magnitude of the response and other factors. Replicates in the number of samples, having replicate gene probes on chips and running replicate chips help increase confidence in the results. Because of the cost of the chips, it is common to pool samples prior to DGE analyses. However, pooling data from replicates on chips is more reliable (29). Although some redundancy is built into several platforms, such as the Affymetrix chip, this does not alleviate the need to assess the variation in expression intensities obtained from different chips and different complex mRNA probes (12). Data processing or elegant protocols cannot substitute for the requirement of multiple independent determinations of the expression intensities (12).

Regardless of the DGE platform and statistics employed, altered expression should be confirmed by an independent method. Common methods to confirm altered mRNA levels are by Northern blotting, standard semiquantitative reverse-transcription PCR (RT-PCR), and quantitative real-time PCR (TaqMan™). Northern blotting has long been considered the gold standard for assessing relative levels of an mRNA when comparing multiple experimental groups. However, it is cumbersome, although relatively easy to perform and the sensitivity of the assays is several orders of magnitude less than the PCR assays.

Standard RT-PCR analysis, although more sensitive than Northern blotting, requires multiple steps following separation of the PCR products on an agarose gel and is subject to contamination. Thus, RT-PCR is inadequate for obtaining quantitative data. Presently, the most widely used method for confirming DGE data is by a real-time quantitative PCR method known as TaqMan (16). In this assay, 2 gene-specific oligonucleotide primers are used to amplify the gene of interest in the presence of a gene-specific, fluorescently labeled probe that binds between the 2 primers. A DNA polymerase cleaves the probe, the reporter dye is separated, and a signal is generated. With real-time monitoring of the PCR reaction by a fluorescence reader the starting copy number of the gene of interest can then be determined based on the cycle at which the PCR signal is first detected. The PCR reaction is usually done in a 96- or more well plate, thus making it possible to test numerous conditions in a single experiment.

Data Normalization

It is important that DGE data generated by microarray analysis be processed to determine whether the putatively differentially expressed genes are really different. Potential sources of variability in microarray experiments include mRNA preparation, transcription, labeling, PCR amplification of clones, variation in spotting (pin geometry, spotting volume, target fixation to membrane), hybridization parameters, slide or membrane inhomogeneities, nonspecific hybridization, nonspecific background, and image analysis (43). The pathologist also helps assure that animal disease, collecting different lobes or portions of tissue, circadian rhythm, and other factors are also considered as potential sources of variability.

Several common normalization approaches help reduce variability (18). One assumption is that the total mass of RNA labeled in each reaction is equal. The spot intensity, although for any particular spot may be higher on one membrane than another should average out over hundreds or thousands of spots. Thus, one can normalize to one. Another approach is to normalize data using a set of housekeeping genes whose expression is assumed to be constant and independent of treatment. Therefore, adjusting intensity levels so that the ratio of the "housekeeping" genes between different membranes is close to 1 can normalize the membranes. Caution should be exercised because experimental data suggests that there really is no universal set of housekeeping genes whose expression remains constant under all treatment conditions. In fact, a recent report states that housekeeping genes tend to be loosely regulated, and often exhibit great (ie, four-fold) differences in expression without functional consequence to the cell (51).

Weakly Expressed Genes and the Fold-Change Problem

One problem encountered using DGE technologies is that low abundance genes, if detectable at all, give low signals. Fold-changes are deceptive for genes that are expressed at close to background levels. For example, if the control intensity for gene A is 1 (arbitrary units), and for the treated is 10, the ratio would be a 10-fold induction. However, if the control value were 4, the fold-change would be a 2.5-fold induction.

Meaningful expression patterns can involve groups of transcripts whose relative abundance changes at levels considerably less than 2-fold but ratio measurements less than 2-fold are often considered to be unreliable in an isolated microarray experiment (22). The authors prefer a statistical cut-off (eg, $p < 0.05$) to a ratio limit (eg, >2-fold). It is particularly valuable to have information on transcripts from genes expressed at low levels because many of the regulatory components of the cell are expressed at low levels (19). Data from in situ hybridizations seem to suggest that the normal variance for many tightly regulated tissue-specific genes is within 20 to 30%. However, there are 2- to 4-fold random fluctuation for many genes in yeast (11, 27).

The number of and complexity of genes that are considered up- and down-regulated, even in a simple experiment, can be daunting. Many gene names give little clue as

#	MLI	Ratio	p-grou	p-value	gene name
476	0.552	24	2	9.28E-29	C06g: stearyl-CoA desaturase 2
573	0.197	10.5	2	1.70E-17	C13m: EST clone RBC121 (rat pancreatic islet cDNA library); similiar to rat ATP
553	3.03	8.54	1	3.97E-14	C12g: steroidogenic acute regulatory protein
592	3.46	8.17	2	3.88E-13	D01d: s-adenosylmethionine synthetase
1038	2.12	7.8	2	1.18E-13	F05b: neural tissue-specific actin filament binding protein (NEURABIN)
562	0.541	7.09	1	1.84E-12	C13b: 3-alpha-hydroxysteroid dehydrogenase (3-alpha-HSD)
421	-1.35	6.89	1	8.07E-13	C03a: aryl hydrocarbon receptor nuclear translocator 2 (ARNT2)
1107	-1.4	6.73	2	1.78E-12	F10a: SM20 protein
794	3.74	6.44	1	1.10E-10	E01j: molecular adapter rGrb
125	0.74	5.09	2	3.80E-09	A09m: LIM, smooth muscle ce
1180	4.71	5.07	2	3.02E-08	G27: glyceraldehyde 3-phosp
134	3	4.7	1	3.24E-08	A10h: lost on transformation 1 protein (LOT1)
188	4.17	4.68	1	8.84E-08	A14f: semaphorin Z (SEMAZ)
789	-0.327	4.54	1	2.97E-08	E01e: insulin receptor substrate 3 (IRS3)
566	-0.911	4.37	2	2.79E-08	C13f: NAD-dependent 15-hydroxyprostaglandin dehydrogenase (HPGD)
446	2.17	4.18	2	1.96E-07	C04l: PFK-L mRNA for liver phosphofructokinase
135	-5.88E-02	4.16	2	1.45E-07	A10i: zinc finger transcription factor homolog CPG20
1085	2.42E-02	4.06	2	2.19E-07	F08g: SMR2
1179	4.07	3.99	2	1.45E-06	G15: hypoxanthine-guanine phosphoribosyltransferase (HPRT)
44	-0.81	3.97	2	2.73E-07	A04b: leucine zipper protein; transcription factor EC (TFEC)
520	1.33	3.94	1	2.36E-07	C10b: cysteine dioxygenase
313	-1.52	3.81	2	7.24E-07	B09e: connexin 36 (CXN36; CX36)
781	0.834	3.79	2	1.26E-06	D14k: ovarian-specific hydroxysteroid 17-beta dehydrogenase 7 (HSD17B7)
336	1.18	3.72	2	1.02E-06	B10n: peroxi
15	3.48	3.65	1	5.61E-06	A02a: 4F2 he
557	-0.547	3.64	1	1.78E-06	C12k: renin-p
778	-1.62	3.6	2	2.47E-06	D14h: osteoprotegerin (OPG); tumor necrosis factor receptor superfamily member
581	1.63	3.56	1	1.91E-06	C14g: D-dopachrome tautomerase
305	0.615	3.49	2	5.50E-06	B08k: potassium channel regulator 1 (KCR1)
75	-0.158	3.46	1	4.86E-06	A06e: nuclear factor I-A1 (NF1-A1)
371	0.136	3.45	2	4.66E-06	B13g: brain-enriched guanylate kinase-associated protein 1 (BEGA1)
62	-0.886	3.45	1	3.10E-06	A05f: homeobox-plus HoxA1 protein [alternatively spliced]
574	3.49E-02	3.44	2	4.63E-06	C13n: GTP cyclohydrolase I
332	3.57	3.24	2	3.73E-05	B10j: translocase of inner mitochondrial membrane 23 homolog (TIM23)
435	2.88	3.23	2	2.13E-05	C0 a: transketolase

Stearyl-CoA Desaturase 2

Transketolase

FIGURE 4.—Partial List (35/1185) of gene expression data for comparison of adipocytes treated with insulin versus no additional insulin. The files are present as output from software developed to statistically compare two group (n = 3 in this case), using local regression (see Crosby et al, 2000). This software is available for download at no cost (ftp://ftp.santafe.edu/pub/kepler/). Key: # = gene number on Clontech Array list (see www.clontech.com, Rat 1.2 nylon array); MLI = mean log intensity, indicating gene signal relative to entire population signal; Ratio = control/treated; Up-group, 2 indicates treated upregulated compared to control; gene name can be checked against membrane location code, see www.clontech.com. Two genes are highlighted as examples used in the diagrammatic speculative interpretation of cellular responses, shown in Figure 5.

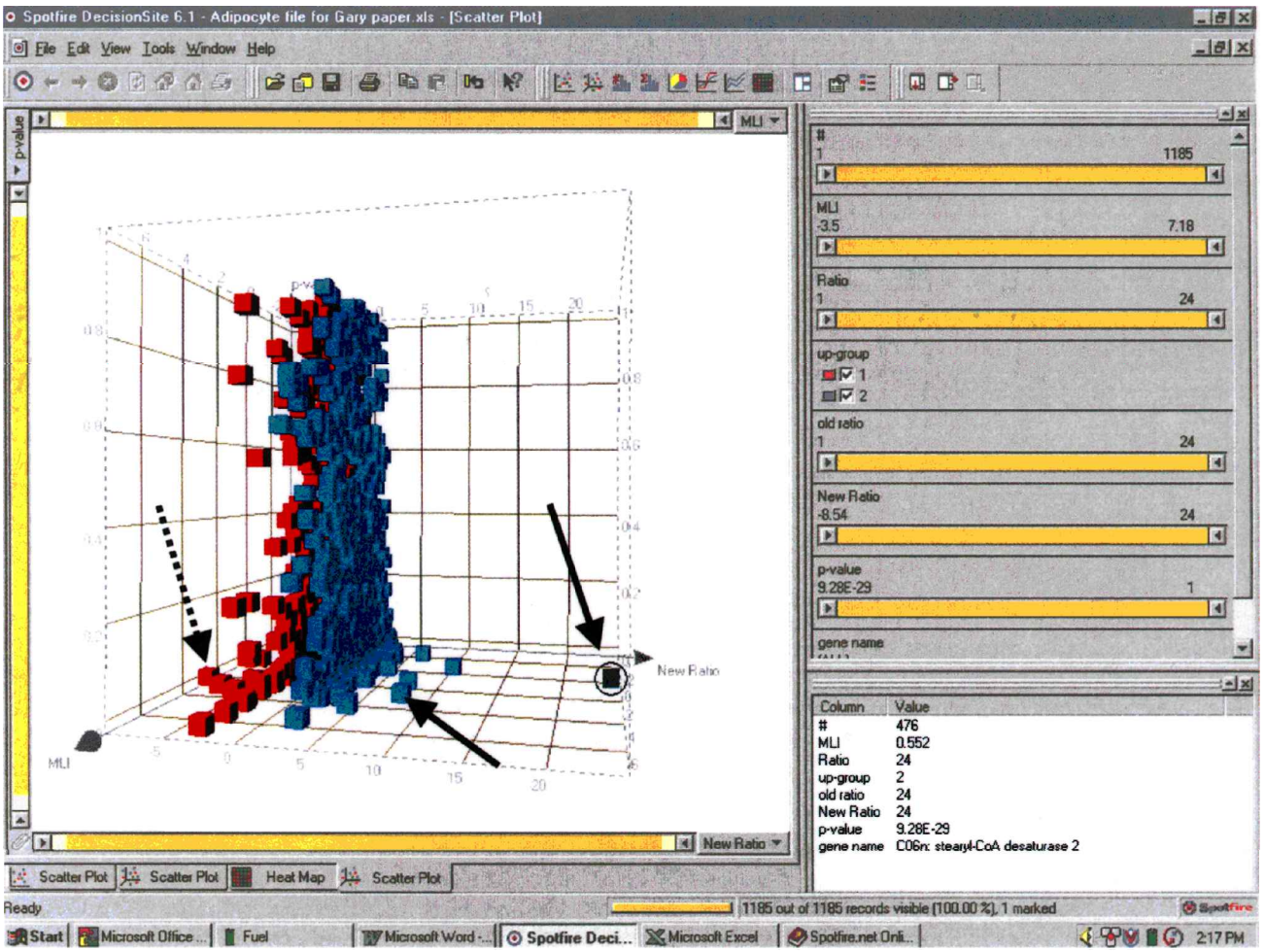


FIGURE 5.—Example of DGE analytical software (Spotfire; insert disclaimer!). Selection of appropriate software for your needs is critical and learning to use the software effectively takes time. This image shows the adipocyte, 1200 gene data set present in a 3-D xy format, with signal strength, p-value, and ratio (control versus treated) on as the respective axes. The short arrow indicates s-adenosymethione synthetase, which is upregulated ~8-fold, has a strong signal, and the ratio is highly significant ($p < .00001$). The long arrow and broken arrows refer to other up (stearyl Co-desaturase) and down (steroidogenic acute regulatory protein, STAR) respectively, which also stand out as responding to treatment. With complex data sets, including dose and time series, such software provides an efficient way to triage the data prior to time-consuming interpretation and costly confirmation.

to their function and many genes have multiple synonyms. For example, the trefoil protein gene (the common red and white clovers are from the trefoil family) would seem to have little relevance for toxicology. However, there are numerous web sites such as (http://www.ncbi.nlm.nih.gov/) that easily provide relevant literature on this gene. This National Library of Medicine (NLM) web site shows that trefoil protein is analogous to human pS2 (BCEI gene) an estrogen-inducible gene expressed in human breast cancers. A brief summary and relevant recent articles on this gene can be obtained under the OMIM link in the NLM web site.

A review of the summary of the up- and down-regulated genes helps organize your thoughts. Before you start looking at your first data set, there are many examples of gene expression data on the Internet (World Wide Web). An excellent example is provided by the following web site: [http://www.sciencemag.org/cgi/content/full/283/5398/83?maxtoshow=&HITS=10&hits=10&RESULTFORMAT=&author1=Iyer&searchid=QID NOT SET&stored search=&FIRSTINDEX=&fdate=10/1/1998&tdate=12/31/2000]. It is strongly recommended that you try your hand at interpreting such data sets before embarking on a career in toxicogenomics. It is also important to be cognizant of the issue of data quality before interpreting DGE data. There are number of stages of data analysis directed towards increasing confidence in both the quality and relevance of DGE data, which are presented in order of increasing importance, as follows:

1. Raw array data, with or without normalization.
2. Statistically confirmed array data, based on multiple replicates, appropriate controls and accepted rules of statistical analysis. Cluster analysis is very helpful in time course studies.
3. Confirmation using second technique, such as Northern analysis or RT PCR (TaqMan™) with a range of experimental designs that address crude signal all the way to measuring mRNA copy number.

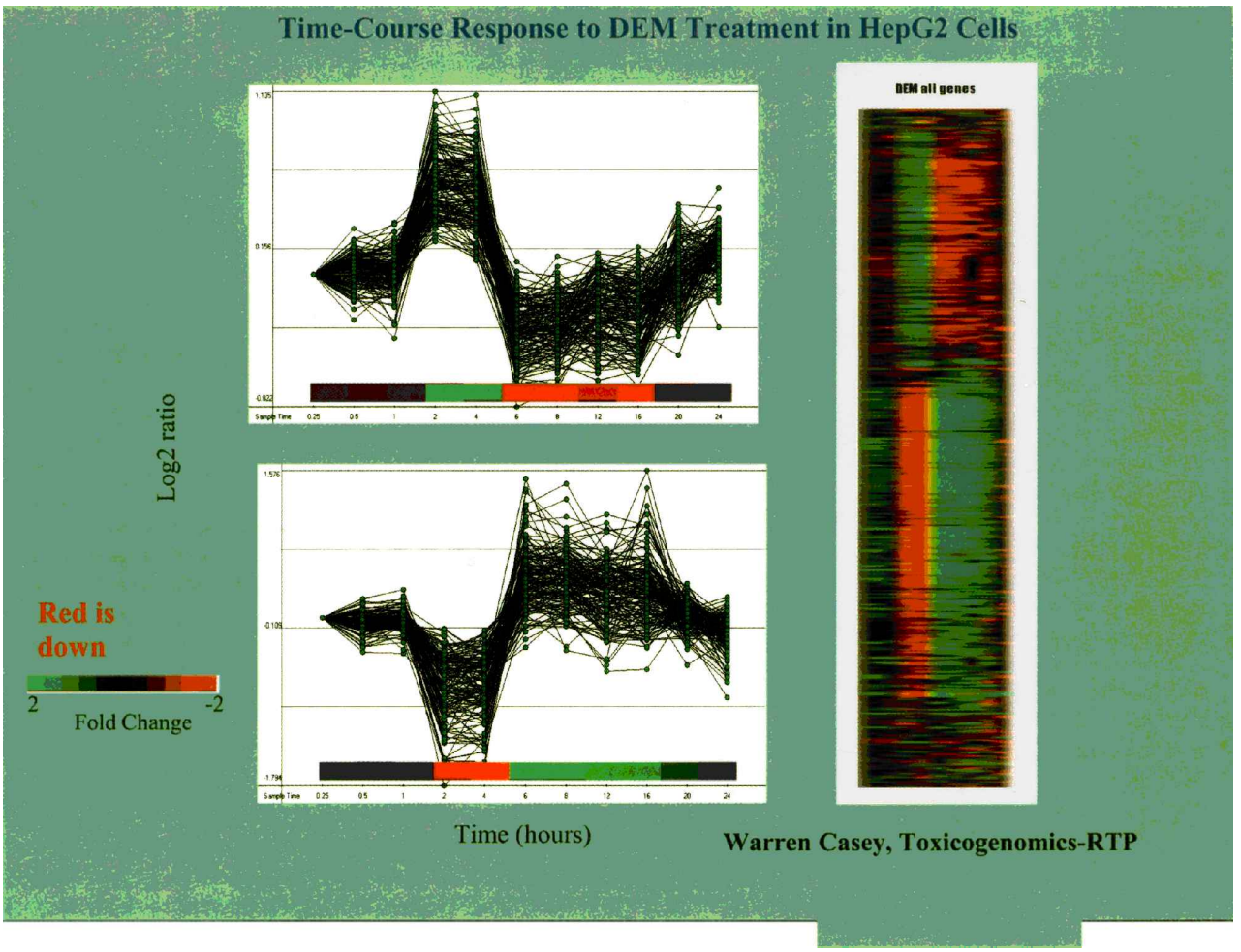


FIGURE 6.—Cluster analysis showing gene expression changes with time in cells following treatment. In this figure, red represents decreased expression and green represents increased mRNA following treatment.

4. Confirmation of an increase in downstream protein.
5. Confirmation of appropriately altered protein activity (eg, enzyme activity). Final confirmation of the relevance of gene expression data is clearly a painstaking but potentially worthwhile process.

AN EXAMPLE OF DGE INTERPRETATION

A small gene expression study of 1,185 genes was undertaken to explore gene expression changes in primary rat adipocytes exposed to a high glucose environment with or without added insulin (Gottschalk et al, unpublished observations). These data are presented here as a simple example of an approach to analyzing data. Following exposure of these cells for 8 hours, the DGE data were developed using Clontech Rat 1.2 arrays (13). A partial list (35/1,185) of gene changes is presented (Figure 4). The first point to be made from Figure 4 is that you will have to become familiar with interpretation of lists. The 2 genes highlighted in this figure demonstrates two important facts:

1. You will encounter many genes about which you know little or nothing and you will have to decide which ones to study. In this case, the pathologist reading these changes had never heard of stearyl CoA desaturase, or the desaturase system. The pathologist is now a little more enlightened.
2. You will encounter old friends, for example transketolase, which is critical in thiamine deficiency (eg, polioencephalomalacia of ruminants). Starting with the genes or biochemical pathways with which you are familiar will help overcome the stress of the numerous genes or pathways with which you are not familiar. The morphology can draw your attention to gene expression components of relevant pathways (apoptosis, cell growth genes in hyperplasia).

The use of gene expression analysis software aids in the investigation of DGE data sets, especially if they are very large or are comprised of numerous exposure concentrations or time points. An example of such software is shown in Figure 5, using the same data set listed in Figure 4. A cluster analysis is shown in Figure 6. Such software permits comparisons of multiple variables in visual formats that are pleasing to pathologists, having a visual bent. If you wish to experiment, download both data (eg, Iyer) and software [http://www.clustan.com/] from the Internet and try these procedures for yourself (strongly recommended).

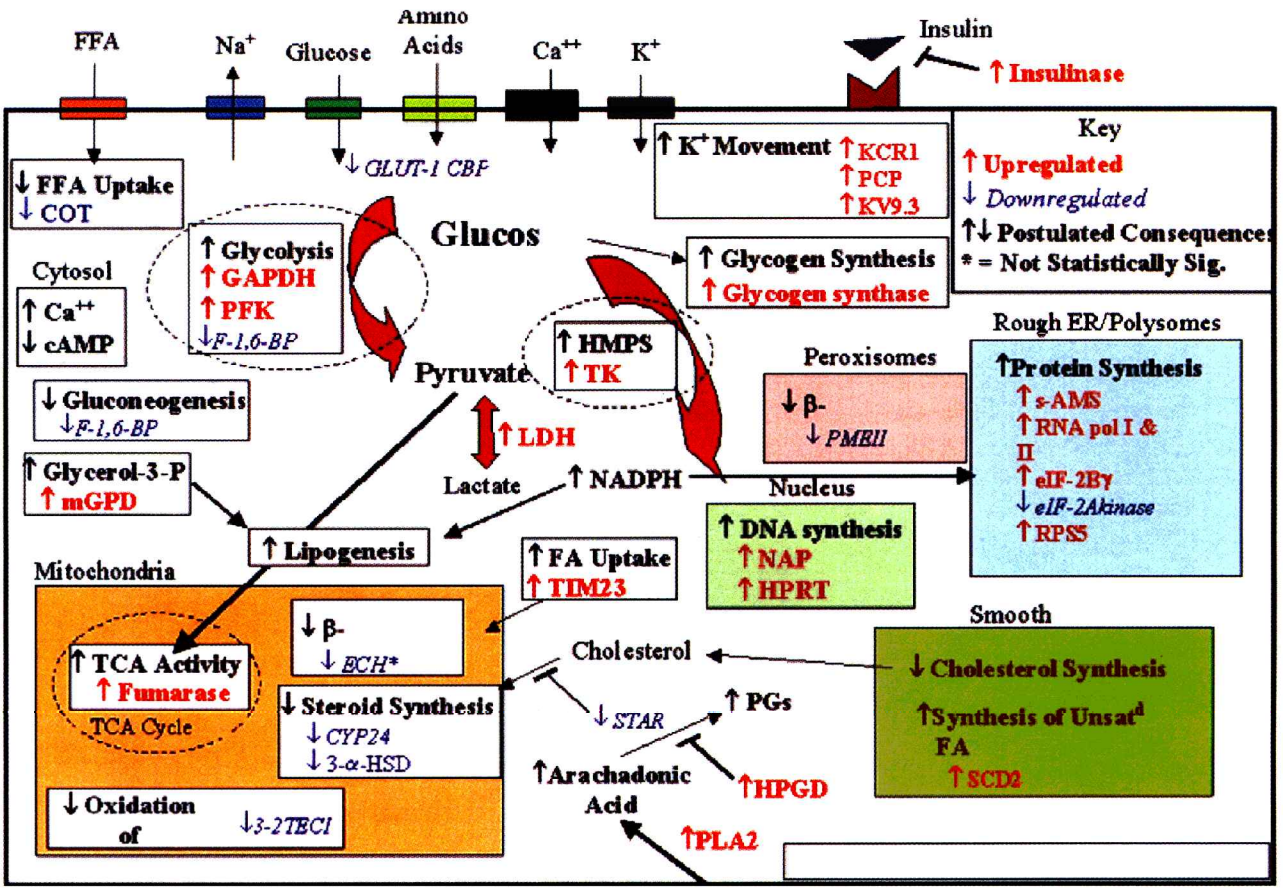


FIGURE 7.—S-ADM = s-adenosylmethionine synthase; RNA pol I & II = RNA polymerase I and II; eIF-2Bγ = eukaryotic initiation factor 2Bγ; RPS5 = 40S ribosomal protein S5; GAPDH = glyceraldehyde 3-phosphate dehydrogenase; KCR1 = potassium channel regulator 1; KV9.3 = shab-related delayed rectifier potassium channel KV9.3; PCP = potassium channel protein; SCD2 = stearyl-CoA desaturase 2; HMPS = hexose monophosphate shunt; TK = Transketolase; SCD2 = Stearyl Co-A Desaturase 2; TIM23 = Translocase of inner mitochondrial membrane 23 homolog; PFK—phosphofructokinase; COT = carnitine octoyltransferase; PMEII—peroxisomal multifunctional enzyme type II; ECH = mitochondrial enoyl-CoA hydratase; 3-2TECI = mitochondrial 3-2trans-trans-enoyl-CoA isomerase; NAP = nucleosome assembly protein; HPRT = hypoxanthine guanine phosphoribosyltransferase.

Finally, you have to examine the data (Figure 4) and determine what they mean for cell or tissue function. In this case, the majority of these responses are consistent with known responses of adipocytes to insulin found in basic biochemistry texts (36). The DGE data were consistent with increased glucose catabolism via glycolysis and Krebs cycle and the hexose monophosphate shunt. The increased use of glucose was combined with increased synthesis of glycogen, triacylglycerol, and unsaturated fatty acids. There was also an increase in protein and DNA synthesis. The DGE was consistent with decreased gluconeogenesis, fatty acid uptake and oxidation, and cholesterol and steroid synthesis. One of the authors (KTM) interpretations of the DGE data is shown diagrammatically (Figure 7). Visual representation of the complex biological processes is helpful in understanding and communicating the results.

Although many of these changes are consistent with insulin treatment, they clearly require further confirmation. The development of such a picture, however, provides guidance to the research team in the form of testable hypotheses. This is a task to which experimental pathologists are ideally suited. Such analyses, combined with pattern recognition approaches, such as those reported by Burczynski et al (8) will provide effective means to extract and apply information in large-scale gene expression data sets. These technologies are also being used for in vivo assessment of tissue responses.

DISCUSSION

The analysis of DGE patterns derived from healthy/control and pathological situations will almost certainly provide a valuable tool in the discovery of therapeutic targets and in the development of diagnostic markers of toxicity. Furthermore, the use of such techniques should place the classification of drug side effects and adverse reactions on a more rational and mechanistic basis and eventually allow patients to receive drugs appropriate for their genotype (7). It is important, however, to point out that transcript profiling is absolutely dependent on "one gene at a time" biochemical toxicology and molecular biology to determine the role of transcriptional responses in altering phenotype (35).

DGE has limited value when used in isolation. Morphological and clinical chemistry tools are helpful in placing the DGE data in perspective. However, the use of DGE arrays is a power tool for getting out side of the box of your thinking; in our experience it always generates surprises and leads to new

understanding of cell and tissue responses, thus proving very enlightening. The identification of molecular fingerprints for specific mechanisms of toxicity is another area that will be a key task for toxicologists and pathologists working in this area for the foreseeable future (8, 35).

Toxicogenomics research is exciting because one is working in a team with diverse areas of expertise where the pathologist can play an important role. The pathologist must learn to function in the toxicogenomics team. The first step is to gain an understanding of the various aspects of the research process. This means attending meetings, participating in study design and learning new terminology. However, the pathologist brings a unique training and background to the team. Most team members have in-depth training in one area, whereas pathologists are generalists with training in biochemistry, nutrition, anatomy, disease processes, and molecular biology.

In the next decade, the postgenomic era, pathology and pathology departments will undergo a series of changes that will redefine the role of the pathologists (3). Training pathologists in pathology informatics, bioinformatics, and genomics is critical to ensure overall leadership of pathology in the postgenomic era. For many of us today, our training will be on the job. Fortunately, the broad in-depth training of the current experimental pathologists puts them in an excellent position to join the genomic revolution with other scientists who are facing the same dilemma of rapidly evolving technologies and an unparalleled amount of new information.

ACKNOWLEDGMENTS

The authors would like to thank Ron Tyler, Chuck Qualls, Warren Casey, Gina Benavides, and many others who have worked on toxicogenomics over the last 4 years at GSK. The views expressed in this article do not necessarily represent official US EPA policy. Mention of trade names or commercial products does not constitute endorsement or recommendations for use. A cooperative US EPA–UNC Chapel Hill Curriculum in Toxicology postdoctoral fellowship supported LMC.

REFERENCES

1. Afshari CA, Nuwaysir EF, Barrett JC (1999). Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. *Cancer Res* 59: 4759–4760.
2. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1994). *Molecular Biology of the Cell.* 3rd ed. Vol. 1. Garland Publishing Inc, New York, p 1294.
3. Becich MJ (2000). The role of the pathologist as tissue refiner and data miner: The impact of functional genomics on the modern pathology laboratory and the critical roles of pathology informatics and bioinformatics. *Mol Diagnosis* 5: 287–299.
4. Bertelsen AH, Velculescu VE (1999). High-throughput gene expression analysis using SAGE. *Drug Discov Today* 3: 152–159.
5. Blanchard K, DiSorbo O, Burris R, Dunn R, Farr S, Stoll R (2000). Toxicogenomics: Understanding the use of microarrays for toxicology studies in vivo. *Toxicologist* 54: 195.
6. Bonaldo MF, Lennon G, Soares MB (1996). Normalization and substraction: two approaches to facilitate gene discovery. *Genome Res* 6: 791–806.
7. Brent R (2000). Genomic biology. *Cell* 1000: 169–183.
8. Burczynski ME, McMillian M, Ciervo J, Li L, Parker JB, Dunn RT, Hicken S, Farr S, Johnson MD (2000). Toxicogenomic-based discrimination of toxic mechanisms in HepG2 human hepatoma cells. *Toxicol Sci* 58: 399–415.
9. Burgess JK, Hazelton RH (2000). New developments in the analysis of gene expression. *Redox Rep* 5: 63–73.
10. Carulli JP, Artinger M, Swain PM, Root CD, Chee L, Tullafallig C, Guerin J, Osborne M, Stein G, Lian J, Lomedico PT (1998). High throughput analysis of differential gene expression. *J Cell Biochem Suppl* 30/31: 286–296.
11. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Morris MS, Davis RW (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2: 65–73.
12. Claverie JM (1999). Computational methods for the indentification of differential and coordinated gene expression. *Hum Molec Gen* 8: 1821–1832.
13. Crosby LM, Hyder KS, DeAngelo AB, Kepler TB, Gaskill B, Benavides GR, Yoon L, Morgan KT (2000). Morphological analysis correlates with gene expression changes in cultured F344 rat mesothelial cells. *Tox Appl Pharmacol* 169: 205–221.
14. Diatchenko L, Lasu Y-F, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD (1996). Suppressive subtractive hybridization: A method for generating differentially regulated of tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci USA* 93: 6025–6030.
15. Fielden MR, Zacharewski TR (2001). Challenges and limitations of gene expression profiling in mechanistic and predictive toxicology. *Tox Sciences* 60: 6–10.
16. Gibson EE, Heid CA, Williams PA (1996). A novel method for real time quantitative PCR. *Genome Res.* 6: 995–1001.
17. Green CD, Simmons JF, Taillon BE, Lewin DA (2001). Open systems; panoramic views of gene expression. *J Immunol Methods* 250: 67–79.
18. Hegde P, Abernathy K, Gay C, Dharap S, Gaspard R, Hughers JF, Snesrud E, Lee N, Quackenbush J (2000). A concise guide to cDNA microarray analysis. *BioTechniques* 29: 548–562.
19. Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95: 717–728.
20. Hubank M, Schatz DG (1994). Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucl Acids Res* 22: 5640–5648.
21. Hubank M, Schatz DG (1999). cDNA representational difference analysis: A sensitive and flexible method for identification of differentially expressed genes. *Methods Enzymol* 303: 325–349.
22. Hughes TR, Marton MJ, Jones AR, Roberts CF, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty A, Simon J, Bard M, Friend SH (2000). Functional discovery via a compendium of expression profiles. *Cell* 102: 109–126.
23. Ishi M, Hashimoto SI, Tsutsumi S, Wada Y, Matsushima K, Kodama T, Aburatani H (2000). Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* 68: 136–143.
24. Kaminski N (2000). Bioinformatics—A users perspective. *Am J Respir Cell Mol Biol* 23: 705–711.
25. Khan J, Bittner ML, Chen Y, Meltzer PS, Trent JM (1999). DNA microarray technology: The anticipated impact on the study of human disease. *Biochem Biophys Acta* 1423: M17–M28.
26. Kim S, Zeller K, Dang CV, Sandgren EP, Lee LA (2001). A strategy to identify differentially expressed genes using representational difference analysis and cDNA arrays. *Anal Biochem* 288: 141–148.
27. Klevecz RR, Kauffman SA, Shymko RM (1984). Cellular clocks and oscillators. *Int Rev Cytol* 86: 97–128.
28. Kornmann B, Preitner N, Rifat D, Fleury-Olela F, Schibler U (2001). Analysis of circadian liver gene expression by ADDER, a highly sensitive method for the display of differentially expressed mRNAs. *Nucl Acids Res* 29: e51.
29. Lee MLT, Kuo FC, Whitmore GA, Sklar J (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci USA* 97: 9834–9839.

30. Liang P, Pardee AB (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 275: 967–971.
31. Lockhart DJ, Winzeler EA (2000). Genomics, gene expression and DNA arrays. *Nature* 405: 827–836.
32. Madden SL, Galella EA, Zhu J, Bertelsen AH, Beaudry GA (1997). SAGE transcript profiles for p-53-dependent growth regulation. *Oncogene* 15: 1079–1085.
33. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, Bassett DE, Hartwell LH, Brown PO, Friend SH (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* 4: 1293–1301.
34. Mathews MB, Sonenberg N, Hershey JWB (2000). Origins and principles of translational control. In: *Translational control of gene expression*, Sonenberg N, Hershey JWB, Mathews, MB (eds). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp 1–31.
35. Morgan KT, Brown HR, Benavides GR, Crosby LM, Sprenger D, Yoon L, Ni H, Easton M, Morgan D, Laskowitz D, Tyler R (2001). Toxicogenomics and human risk assessment. *Health Environ Risk Assess* (in press).
36. Murray RK, Granner DK, Mayes PA, Rodwell VW (2000). *Harper's Biochemistry*. 25th ed, Vol 1, McGraw-Hill, New York.
37. Notterman DA, Uri A, Sierk AJ, Levine AJ (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissues examined by oligonucleotide arrays. *Cancer Res* 61: 3124–3130.
38. Pennie WD, Tugwood JD, Oliver GJA, Kimber I (2000). The principles and practices of toxicogenomics: Applications and opportunities. *Toxicol Sci* 54: 277–283.
39. Prashar Y, Weissman SM (1996). Analysis of differential gene expression by display of 3′ end restriction fragments of cDNAs. *Proc Nat Acad Sci USA* 93: 659–663.
40. Raitio M, Lindroos K, Laukkanen M, Pastinen T, Sistonen P, Sajantila A, Syvanen AC (2001). Y-chromosomal SNPs in Finno-Ugric-speaking populations analyzed by minisequencing on microarrays. *Genome Res* 11: 471–482.
41. Rashidi HH, Buehler LK (2000). *Bioinformatics Basics*. 2000, Boca Raton, FL: CRC Press, p 185.
42. Schena M, Davis RW (1999). Genes, genomes, and chips. In: *DNA microarrays: A practical approach*, Schena M (ed). Oxford University Press, Oxford, UK, pp 1–16.
43. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, ILehrach H, Herzel H (2000). Normalization strategies for cDNA microarrays. *Nucl Acids Res* 28: e47.
44. Shimkets RA, Lowe DG, Tai JT, Sehl P, Jin H, Yang R, Perdki PF, Rothberg BE, Murtha MT, Roth ME, Shenoy SG, Windemuth A, Simpson JW, Simmons JF, Daley MP, Gold SA, McKenna MP, Hillan K, Went GT, Rothberg JM (1999). Gene expression analysis by transcript profiling coupled to a gene database query. *Nat Biotechnol* 17: 798–803.
45. Smith CG (1992). *The Process of Drug Discovery and Development*. Boca Raton, CRC Press, p 192.
46. Sugita M, Sugiura M (1994). The existence of eukaryotic ribonucleoprotein consensus sequence-type RNA-binding proteins in a prokaryote, Synechococcus 6301. *Nucl Acids Res* 22: 25–31.
47. Sutcliffe JG, Foye PE, Erlander MG, Hillbush BS, Bodzin LT, Durham JT, Hasel KW (2000). TOGA: An automated parsing technology for analyzing expression of nearly all genes. *Proc Natl Acad Sci USA* 97: 1976–1981.
48. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995). Serial analysis of gene expression. *Science* 270: 484–487.
49. Vidal M (2001). A biological atlas of functional maps. *Cell* 104: 333–339.
50. Wang A, Pierce A, Judson-Kremer K, Gaddis S, Aldaz CM, Johnson DG, MacLeod MC (1999). Rapid analysis of gene expression (RAGE) facilitates universal expression profiling. *Nucl Acids Res.* 27: 4609–4618.
51. Warrington JA, Nair A, Mahadevappa M, Tsygansdaya M (2000). Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* 2: 143–147.
52. Zar JH (1999). *Biostatistical Analysis*, 4th ed. Upper Saddle River, NJ, Prentice-Hall, p 359.
53. Zhang JS, Duncan EL, Chang AC, Reddel RR (1998). Differential display of mRNA. *Mol Biotechnol* 10: 155–165.
54. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW (1997). Gene expression profiles in normal and cancer cells. *Science* 276: 1268–1272.

Appendix

Brief Overview of DGE Technology and Data Analysis

The cost, percent coverage of the genome, and the sensitivity of the method need to be considered in selecting a platform for gene expression. Sensitivity is important, because 90 to 95% of mRNA transcripts in a cell are present at 5 or fewer copies per cell (Table 1). These low copy mRNAs make up only 30 to 50% of the total cellular mRNA mass (2, 6, 54). These genes are important because many of the most tightly regulated regulatory components of cells are expressed at low levels (19). In contrast, cellular maintenance or housekeeping gene transcripts are generally present in moderate levels of about 10 to 50 copies per cell (51).

The methods for surveying the differential gene expression (DGE) of the transcriptome technologies can be grouped under 2 broad subdivisions. A closed DGE technology is when the genes of inquiry are predetermined by their inclusion on the slide (or chip). The 2 most common closed systems are microarray hybridization technologies and quantitative polymerase chain reaction (qPCR or TaqMan). Open architecture systems, in contrast, require no a priori comprehensive knowledge of the transcriptome. Both open and closed systems have merits and drawbacks, but are complementary.

Closed Architecture Systems

Closed DGE discovery platforms are commercially available and have relatively high throughput. Disadvantages include variable coverage of the transcriptome and, in some cases, the relatively high cost. However, these nucleic acid microarrays, or DNA chips may revolutionize genetics in the same way that silicone chips revolutionized the computer industry (31). DNA microarrays contain hundreds to thousands of gene-specific sections of DNA (probes) and are generated by 1 of 2 basic methods. cDNA arrays are produced by depositing 200 to 1,000 base pair (bp) sections of DNA (the probes) derived from PCR products or plasmids onto a sold support, such as nylon membranes or glass or plastic slides. Typically, each spot contains 1 to 10 nanograms of DNA, ensuring that saturation does not occur during hybridization with the heterogeneous population of labeled cDNAs from treated and control cells. The cDNAs are synthesized from the total mRNAs collected from animals or cell cultures with different dyes used to distinguish control and treated cells. Oligonucleotide arrays are synthesized,

Table 1.—The populations of mRNA molecules in a typical mammalian cell.

	Copies per cell of each mRNA sequence	Number of different mRNA sequences in each class	Total number of mRNA molecules in each class
Abundant class	12,000	4	48,000
Intermediate class	300	500	150,000
Scarce class	15	11,000	165,000

(From Reference 2, p 369.

base-by-base, on a glass slide in situ using photolithography. The oligonucleotides are typically 20 to 25 bp in length, and each spot contains over a million copies. In both cases, probes are designed from sequence located near the 3′ end of the gene, which is the most variable area of the gene. This is critical for the specificity of the gene. Typically for oligonucleotide arrays, multiple probes per gene are placed on the array, designed to complement regions from several exons of the gene. In addition, multiple negative control, consisting of oligonucleotides with a 1 bp mutation are also synthesized. Both types of arrays are hybridized, usually overnight, with fluorescently- or radioactively-labeled heterogeneous target prepared from 1st strand cDNA synthesis of the total mRNA population for a particular treatment group. The arrays are scanned (fluorescently labeled) or are exposed to phosphorimager screens and then scanned (radioactive label). Computer software packages and statistical testing are then used to identify genes that are differentially expressed.

Advantages of the nylon membrane arrays are affordability, sensitivity, and the ability to reprobe them after stripping. Disadvantages include low spot density and the requirement that each labeled target population must be hybridized to a separate membrane. Advantages of the glass slide based cDNA arrays are high spot density coupled with the ability to hybridize both a control and an experimental sample on the same slide, thus removing the slide-to-slide variability encountered while using nylon arrays. Fluor flips or reversing dyes between controls and treated are needed to eliminate differences due to dye incorporation. Also, glass slides are relatively economical if developed in-house. Advantages of the oligonucleotide arrays include the incorporation of negative controls for each gene and high spotting density. The major disadvantage is the high cost of the commercially available gene chips but the costs are dropping rapidly.

Open Architecture Systems

Open DGE discovery platforms present the advantage of whole-transcriptome coverage. Disadvantages include the relatively low throughput and the high degree of difficulty in carrying out the experimental manipulations. Several open platforms exist including differential display (17), representational difference analysis (14, 20), serial analysis of gene expression (48), GeneCalling (44), total gene expression analysis (47), rapid analysis of gene expression (50), and restriction enzyme analysis of differentially expressed sequences (39). Several are available commercially, either as prepackaged reagents for in-house use, or as a contract service performed by the vendor.

Differential Display

Differential display (30) is the progenitor technology of all current major technologies for transcriptome profiling. Differential display of mRNA is a technique where the mRNA species expressed by cells are reverse transcribed and then amplified by many separate polymerase chain reactions (PCRs) (17). PCR primers and conditions are chosen so that any given reaction yields a limited number of amplified cDNA fragments, permitting their visualization as discrete bands following gel electrophoresis (53). Included in the PCR reaction is a radioactively labeled nucleotide, which allows visualization of the PCR products after electrophoretic separation. The radioactive bands unique to a particular treatment group are then excised from the gel and are re-amplified, this time in absence of the radionucleotide, subcloned, and sequenced. Advantages of differential display include the relatively low cost compared to other DGE technologies and the ability to compare multiple groups, in parallel, simultaneously. A particular problem with differential display is a high rate of false positives. This platform is commercially available from GeneHunter Corporation.

Representational Difference Analysis (RDA)

A subtractive hybridization-based approach to DGE profiling was reported in 1994 (20) and later modified to significantly reduce false positives (14). To perform RDA, RNA from 2 different cell populations, a treated and a control, is converted into cDNA and then fragmented using restriction enzymes. After cleanup, different oligonucleotide linkers are ligated to the ends of the 2 different cDNA populations, and the cDNAs are amplified by PCR. The resulting amplicon pools are labeled with either radioactivity or with biotin. Using an excess of biotin cDNA, the 2 cDNA populations are mixed together, denatured, and are allowed to reanneal. Capture of the biotin label removes all of the cDNA species that are present in both control and treated, leaving only those species unique to the treated sample. Usually, several iterations of this process are necessary to isolate most of the unique cDNAs. A modification can be used to identify down regulated genes. Although this method does not produce quantitative data, a wide range of expression differences (2- to 80-fold) can be detected as verified by Northern blots with good sensitivity. Transcripts representing only 1% of a cell's mRNA population can be identified (46). The drawbacks are that the technique is technically demanding and is prone to high rates of false positives and negatives if not properly performed (21). Recently, a strategy has been reported that selects genes expressed at low abundances (26). A modified version of this platform is commercially available from Clontech.

Serial Analysis of Gene Expression (SAGE)

SAGE is a sequence-based DGE technology that identifies differentially expressed genes and, unlike other open architecture methods, quantifies the level of expression (48). Double-stranded cDNA is synthesized using a biotin-labeled oligo d(T) primer. After digestion of the cDNA with a frequent-cutting restriction endonuclease, their biotin tags capture the 3′-most end of the cDNA fragments. Specific adapters are ligated to the cDNA fragments and the cDNAs are again fragmented with a different restriction enzyme to fragments consisting of a short section of linker DNA attached to a 11-base pair portion of the differentially expressed cDNA. Linker DNA is removed and 25 to 50 of the resulting specific sequence tags are concatenated and sequenced. The result of this procedure is a library of clones where each clone includes a short, unique tags for 20 or more genes and it is possible to generate data points for thousands of genes (10). The procedure is quantitative in that the number of times a given tag shows up on a sequencing run allows enumeration of the number of copies of a particular mRNA species. (48). The coverage and sensitivity of SAGE is dependent upon the number of concatener clones sequenced.

Computer simulations predict the likelihood of detecting a transcript present at three copies per cell, assuming a total of 300,000 cellular transcripts, to be 92% if a total of 300,000 tags, or 10,000 to 15,000 clones are sequenced (4, 32). The primary advantage of SAGE over other technologies is that it is quantitative. Disadvantages include the complexity of the experimental manipulations required and the number of tags that must be sequenced in order to get significant coverage of the transcriptome. In a recent report, genes ranked high by the SAGE generally showed high-intensity scores in GeneChip analysis, although there was relatively poor correlation among genes with lower fold changes (23). This platform is commercially available from Invitrogen.

Other Open Systems

GeneCalling (44) is a DGE platform that relies more extensively on the restriction enzyme digest concept. GeneCalling covers approximately 95% of an expressed genome. This technology is available as a contract research service from Curagen. Total gene expression analysis (TOGA) is another DGE platform that is restriction enzyme-based (47). A sensitivity of at least 1/100,000 is claimed and gene coverage of 60% is reported. However, performing up to four iterations with different restriction enzymes, up to 99% coverage is possible. Rapid analysis of gene expression (RAGE) similar to SAGE, can be used in a targeted manner to search for known genes (50). In this procedure, unique, gene-specific tags averaging ~128 bp are generated from each cDNA by poly(A) selection and digestion with two restriction endonucleases, ligated to common primer binding sites, amplified by PCR, and are electrophoretically separated. The procedure is relatively quantitative in that the intensity of the amplimer band on the gel is a relative measure of the frequency of the corresponding mRNA in the total population of mRNAs. Other DGE technologies include restriction enzyme analysis of differentially expressed sequences (39) and amplification of double-stranded cDNA end restriction fragments (28). These newer technologies are reportedly very sensitive but their utility for research teams is unproven.